Group assignment – Data Mining

Data Mining

Post-graduation in Business Analytics and Business Intelligence

Group -6

Contents

[1. Objective 2](#_Toc33909149)

[2. Assumptions 2](#_Toc33909150)

[3. Thera Bank - Loan Purchase Modelling 2](#_Toc33909151)

[3.1 EDA - Basic data summary, Univariate, Bivariate analysis, graphs 3](#_Toc33909153)

[3.2 Applying Supervised Machine Learning Techniques (Test & Train) 11](#_Toc33909154)

[3.3 Applying CART – Full tree 12](#_Toc33909155)

[3.3 Interpret the CART model output (pruning, remarks on pruning, plot the pruned tree) 13](#_Toc33909156)

[3.3 Applying Random Forests (plot the tree) 14](#_Toc33909157)

[3.3 Interpret the RF model output (with remarks, making it meaningful for everybody) 15](#_Toc33909158)

[4. Various Model Performance Measures (Test & Train): Confusion Matrix 16](#_Toc33909159)

[5. Remarks on Model validation exercise (Which model performed the best) 16](#_Toc33909160)

# Objective

For the given problems we must do the EDA, find FA/PCA and come up with best fit models.

# Assumptions

The Given data is a sample of the population.

# Thera Bank - Loan Purchase Modelling

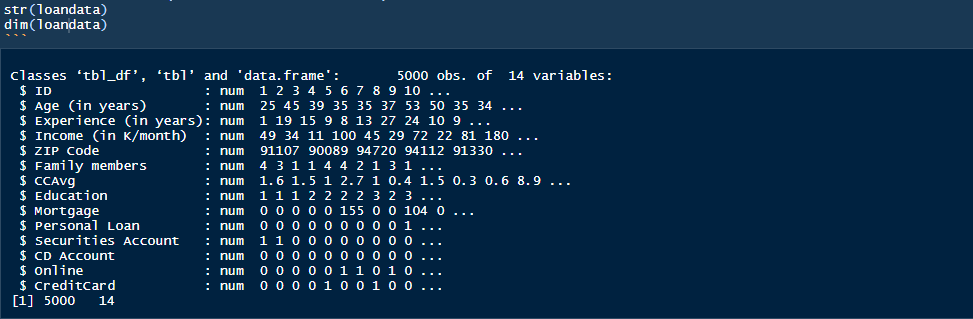
## 

## This case is about a bank (Thera Bank) which has a growing customer base. Majority of these customers are liability customers (depositors) with varying size of deposits. The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans. In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with a minimal budget. The department wants to build a model that will help them identify the potential customers who have a higher probability of purchasing the loan. This will increase the success ratio while at the same time reduce the cost of the campaign. The dataset has data on 5000 customers. The data include customer demographic information (age, income, etc.), the customer’s relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign.

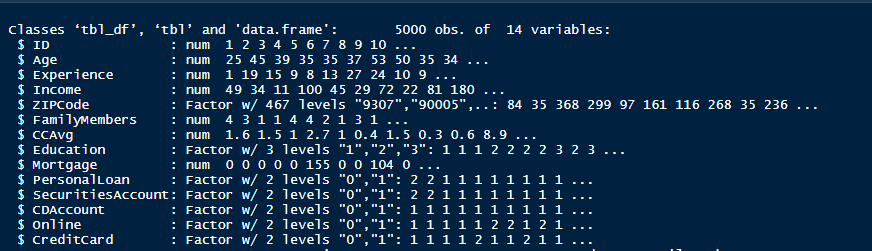


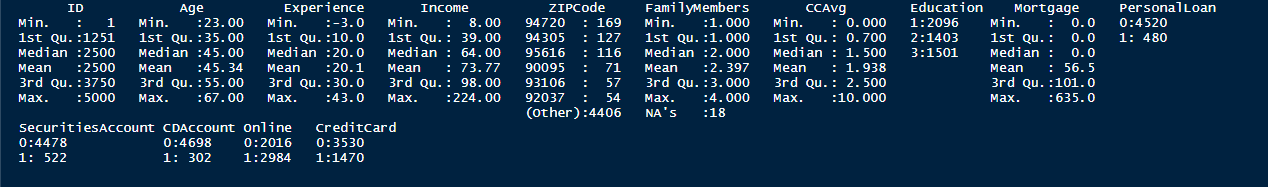
## 3.1 EDA - Basic data summary, Univariate, Bivariate analysis, graphs

**Basic summary of Data**

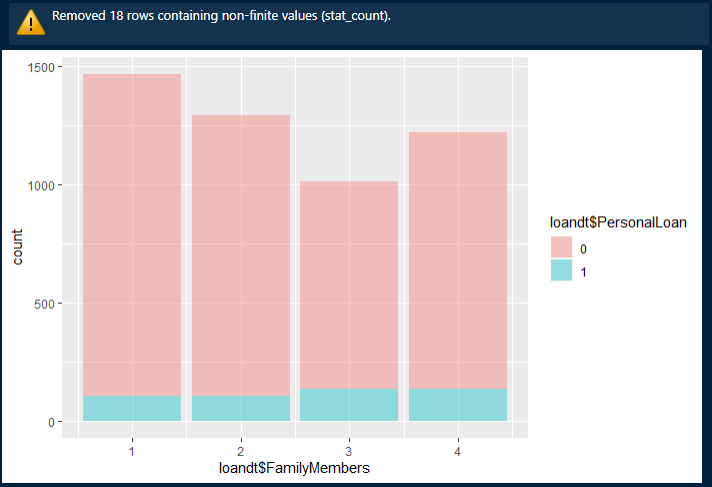


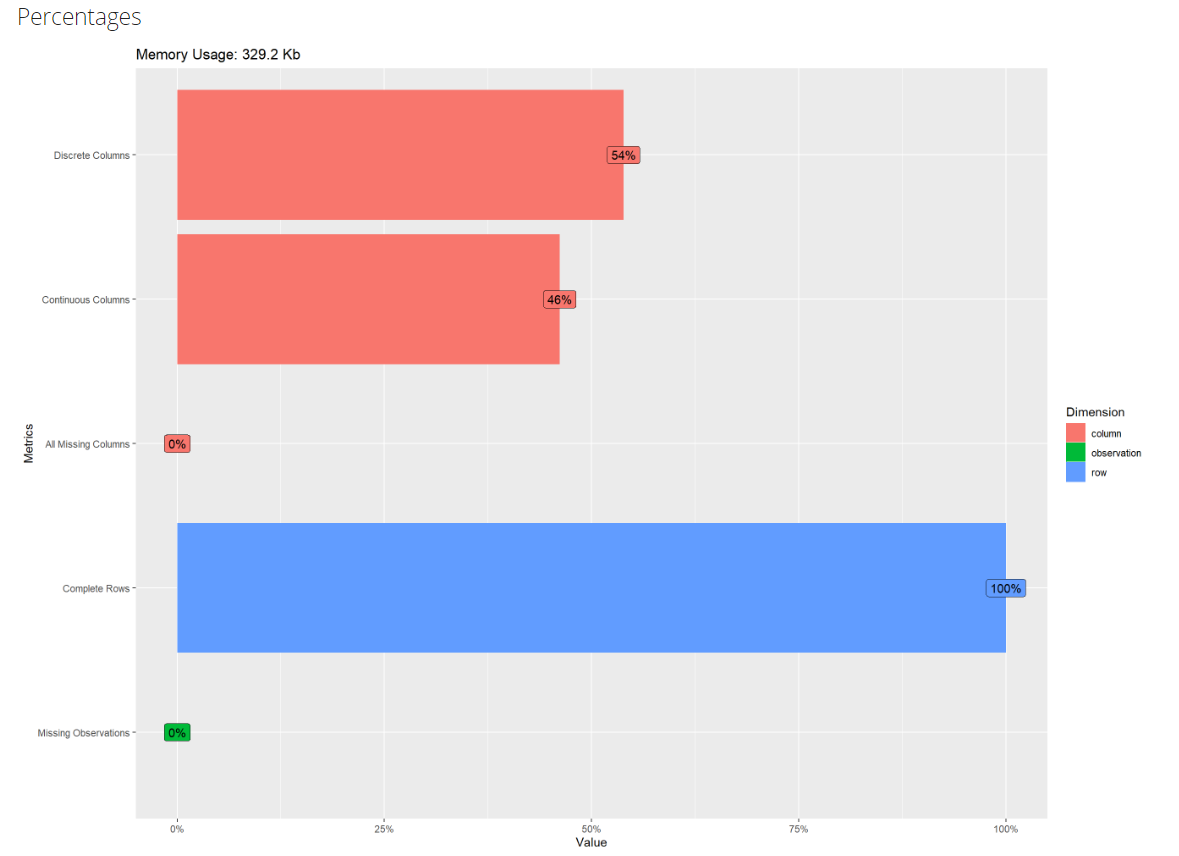
Structure and Summary after converting Zip Code , Education, Personal loan, Securities Account, CD Account, Online and CreditCard to factor

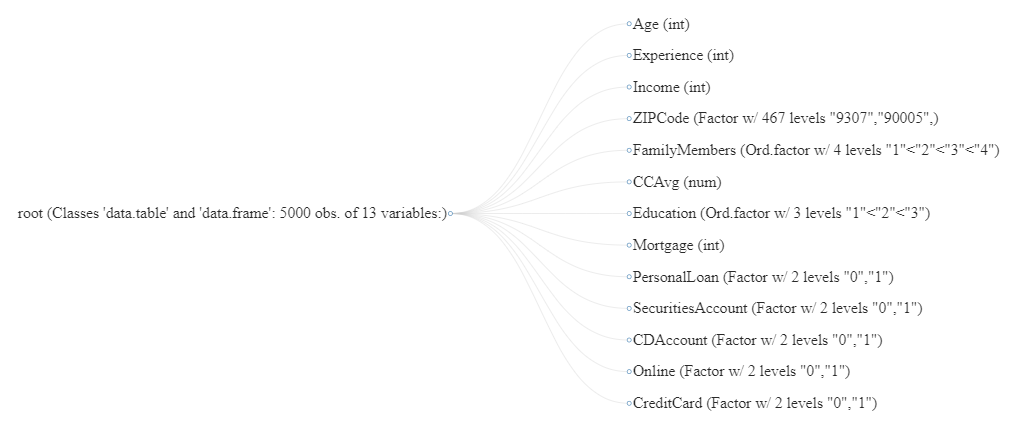




1. Our field of interest is PersonalLoan field and last year only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign so there is data imbalance in our dataset.
2. 75% of people that were targeted were below Age 55 and have below 30 Years of professional experience.
3. It is a good mix of Undergrad, Graduate and Advanced/Professional. Converted Education into ordered factors.
4. We do not need column ID and Zip Code so will remove it from dataset, experience columns has negative values so we will take abs values assuming it was data entry error.
5. Looking at effect on Family members’ size, it has no impact on accepted personal loan. So we can impute the missing values for 18 rows to 1 and converting it to ordered factors.

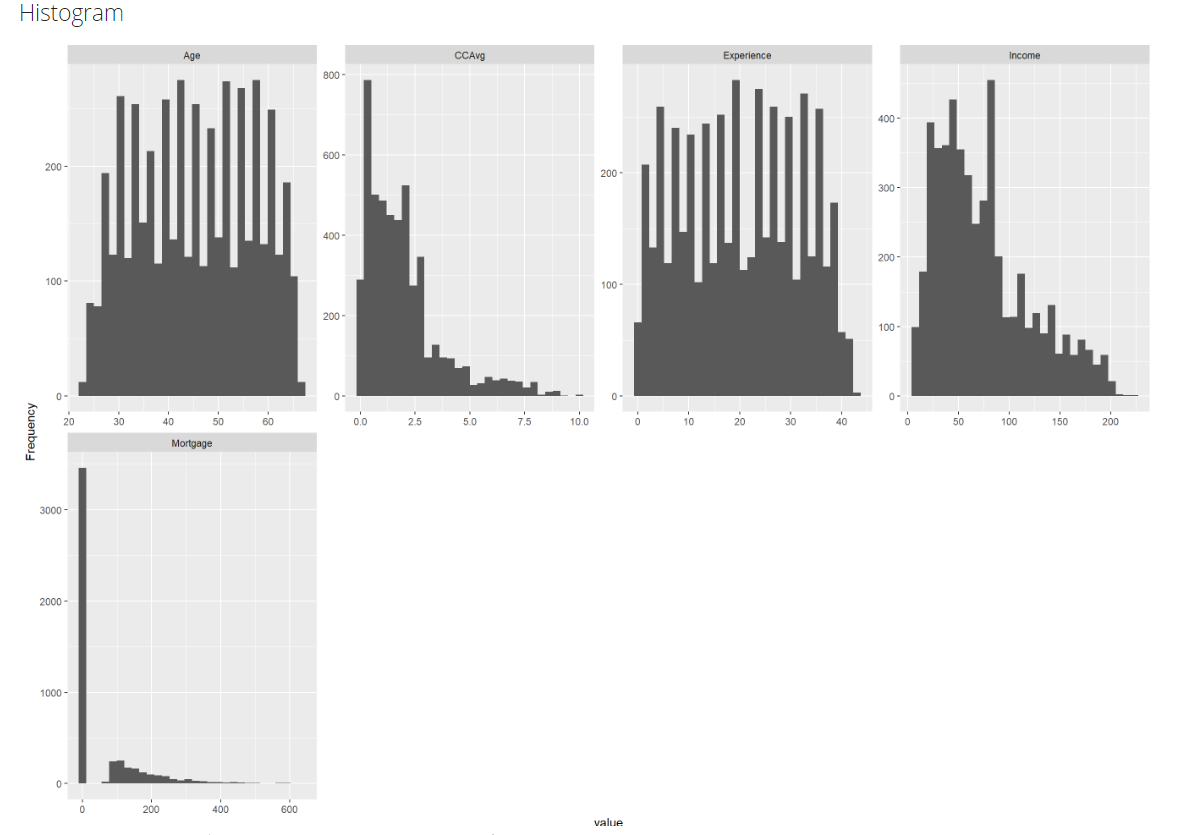


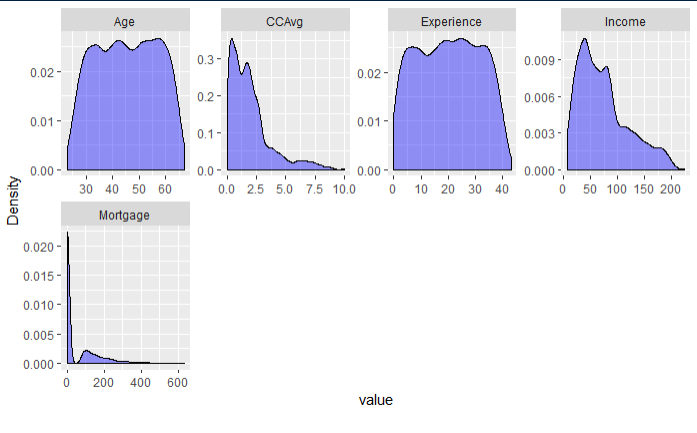
**Basic Data Profile after cleansing and imputing missing values**

****

**Univariate Analysis**

Plotting Histogram and Density for all numerical x variables.

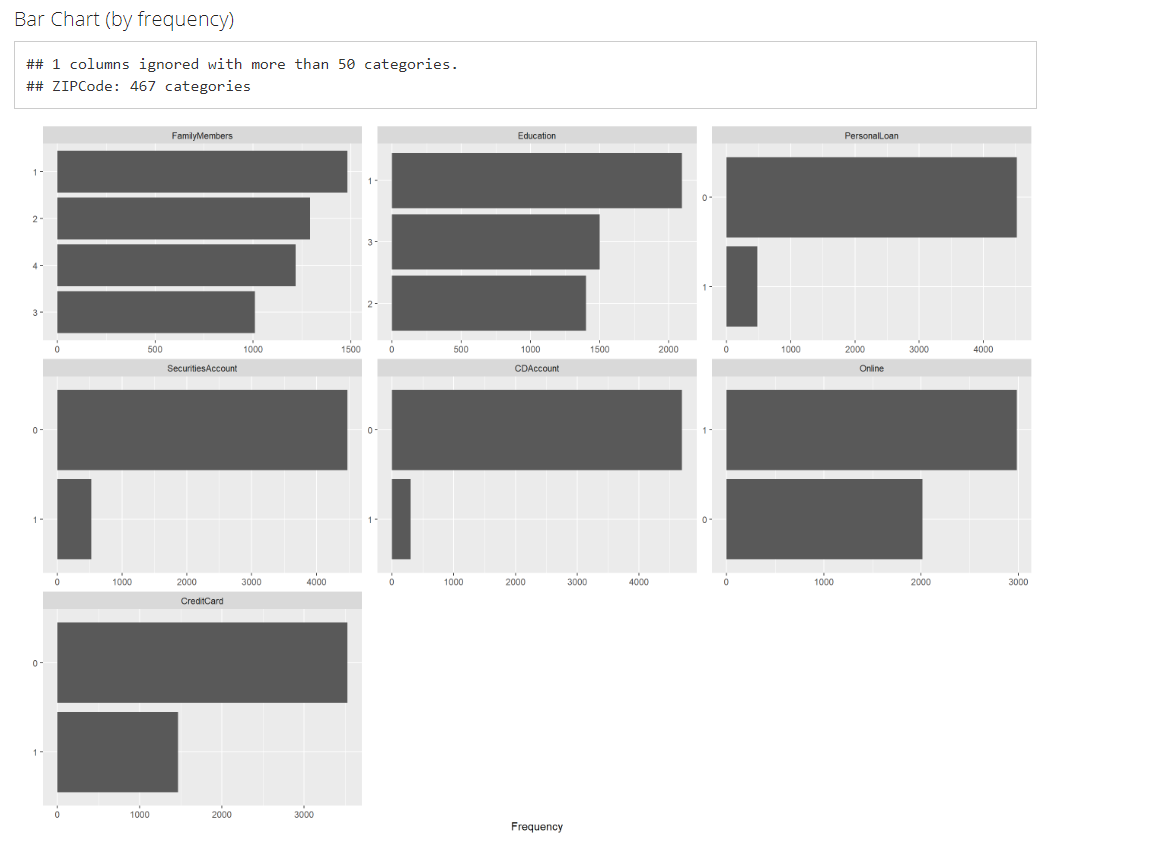
****

****

Observation

* Data is normally distributed by Age and Experience. Data is left skewed for Credit card average spend, Income and Mortgage.
* Majority of customers spend 5K per month using credit card.
* Most of the customers age fall in the age range of 30 to 60 yrs and their experience fall in the range of 5 to 35 years and most earn an income between 10K to 100K.

Bar Chart (by frequency) for Categorical Variables

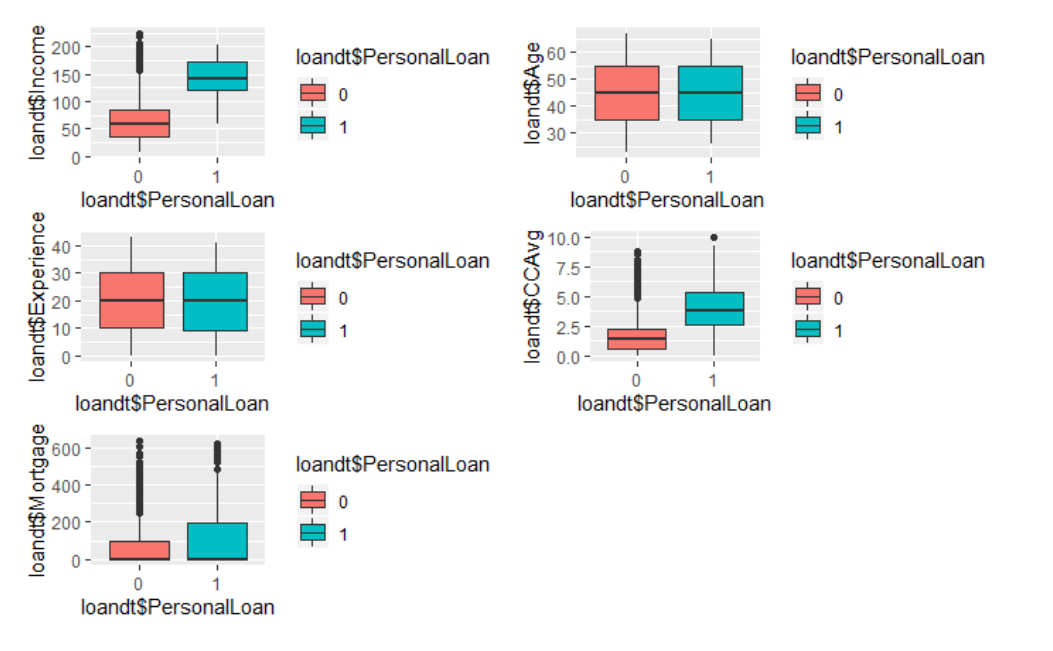


Observation

* Data is skewed when we look at the fields CD Account, Securities Account and credit card issued by the bank.
* Good distribution across Family members, Education and Online banking facilities.

**Bivariate Analysis**

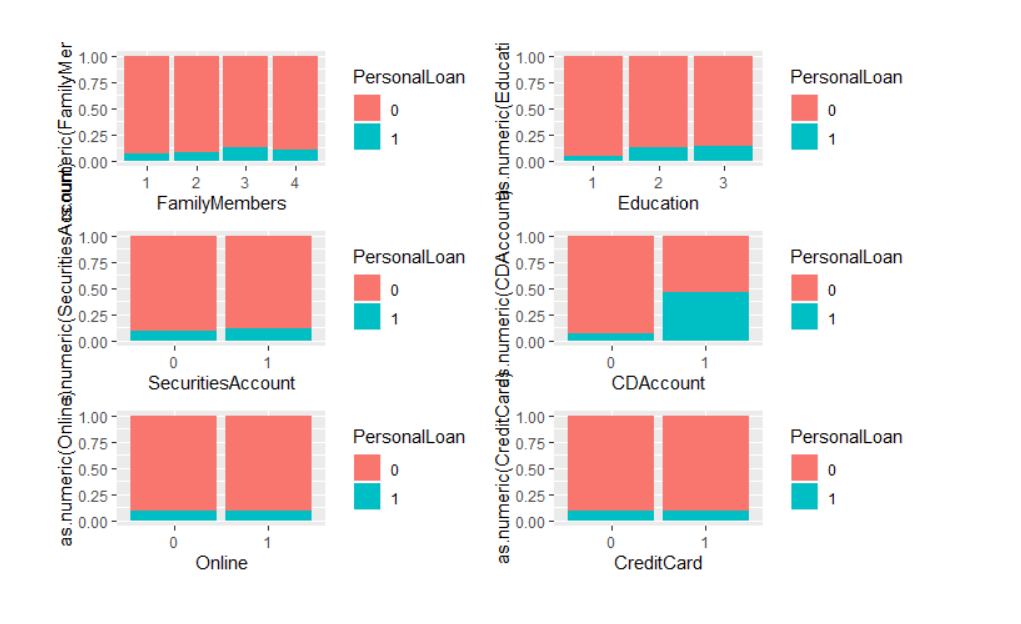
Box plotting Personal Loan against numerical variables.

****

Observation

* Credit card monthly spend, Income and Mortgage seems to be directly affecting personal loan decision.
* Lots of non personal loan (Class 0) takers are present as outliers in Credit Card spend, Mortgage and Income. These could be a very good Target Class.
* Age and Experience have no bearing on the personal loan acceptance.

Stacked Bar plot of Personal Loan for categorical variables.



Observation

* Only CD Account holders have some affect on personal loan acceptance.
* Data is mostly skewed towards non personal loan takers and other categorical variables are not impacting a lot.
* Education class Advanced/Professional has slightly better conversion rate in terms of acceptance.
* Family sizes of 3 are highest acceptors of loan offer.

Creating scatter plot for numerical variables impacting field of interest



Observation

* Customers with higher income have more mortgages and credit card spend.
* Acceptance is higher for customers with more income.
* People having income between 20 K to 90K have no Personal loans and moderate Credit Card spending 2.5K.
* Customers with high credit card spend and income has good potential of conversion and can be targeted.
* Customers with no mortgages have accepted personal loans more; they can be a good segment of customers for targeting.

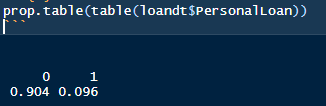
## 3.2 Applying Supervised Machine Learning Techniques (Test & Train)

Positive Class is Personal loan Accepted – “1”, which is ~10% of the dataset so there is significant class imbalance.

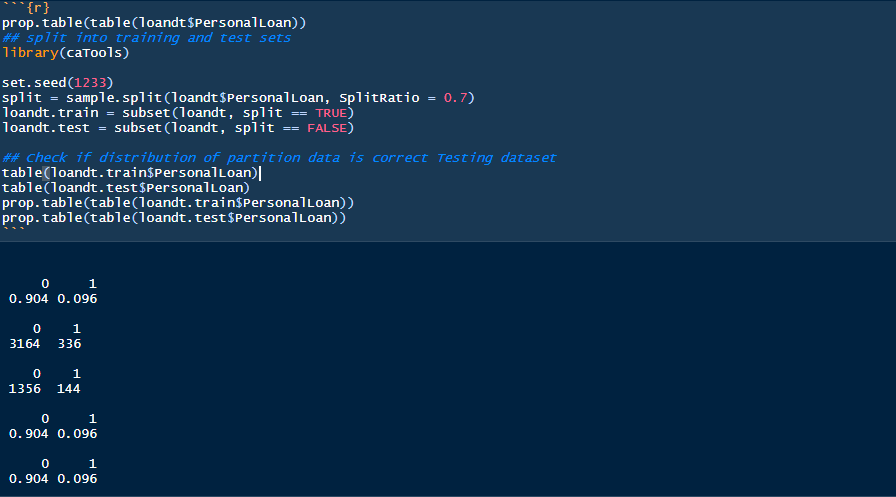
FP – Actually will not accept loan offer but predicted as they will accept

FN – Actually will be taking the offer but predicted as they will not accept the offer.

For our use case FN will be very costly because we will be losing out on potential customers but FP is something that we can live with as it will only lead to additional promotion cost. Ideally we should not even have a single False Negative scenario.

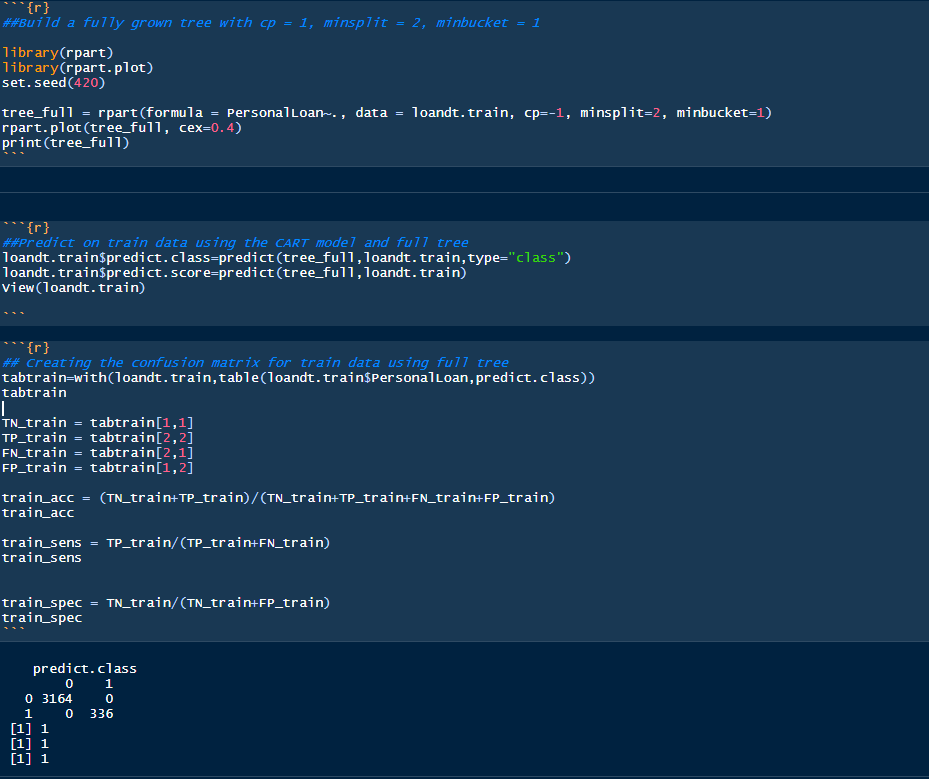


Splitting the data into train and test based on 70/30 ratio fixing the percentage of personal loan same as initial data.

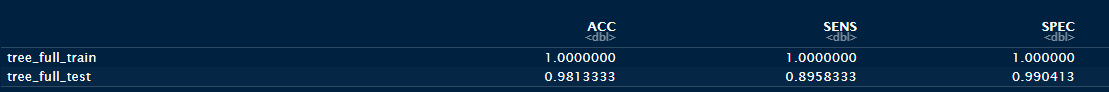


## 3.3 Applying CART – Full tree

Running a cart with full tree for prediction and creating the confusion matrix on train data. As we see below, the model is over fitting and there are no false positives or false negatives.



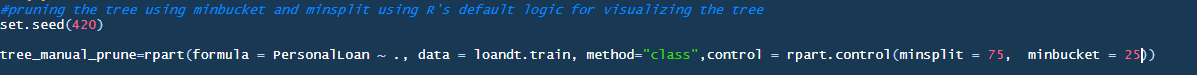
Once we run the model on test data and consolidate the results, we observe that model overfits especially in terms of Sensitivity which deals with false negative and is very costly in our use case.



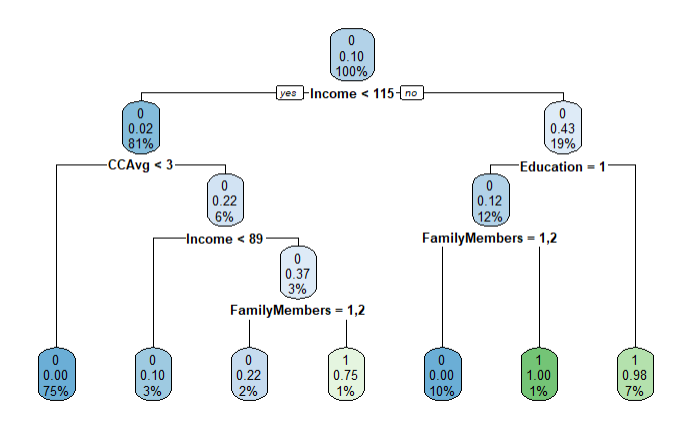
## 3.3 Interpret the CART model output (pruning, remarks on pruning, plot the pruned tree)

Classification trees use recursive partitioning algorithms to learn and grow on data, pruning will be done to limit the recursive based on criteria.

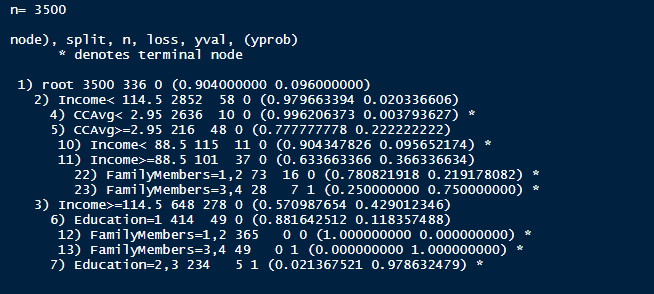
We will first do manual pruning using R’s default logic of minsplit as 1-3% of train data and minbucket as ‘minsplit/3’.



**Plotting the pruned tree**



**Printing the pruned tree**

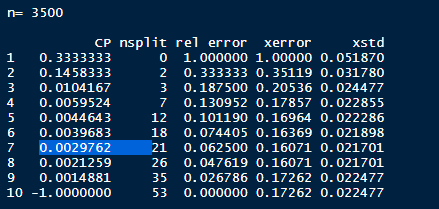
**** Observation

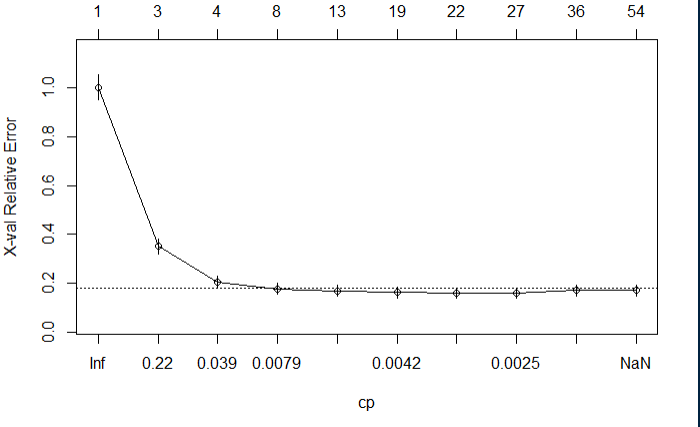
* Income, CC Avg, Family Members and Education are important predictors on which data is split by pruned tree algorithm.
* First split happens on whether Income is less than or greater than $ 114K.
* Second split happens on whether the monthly credit card is less than or greater than 2.95K.
* Also using the manual prune tree we observe that Sensitivity of the model has actually decreased to ~83%.



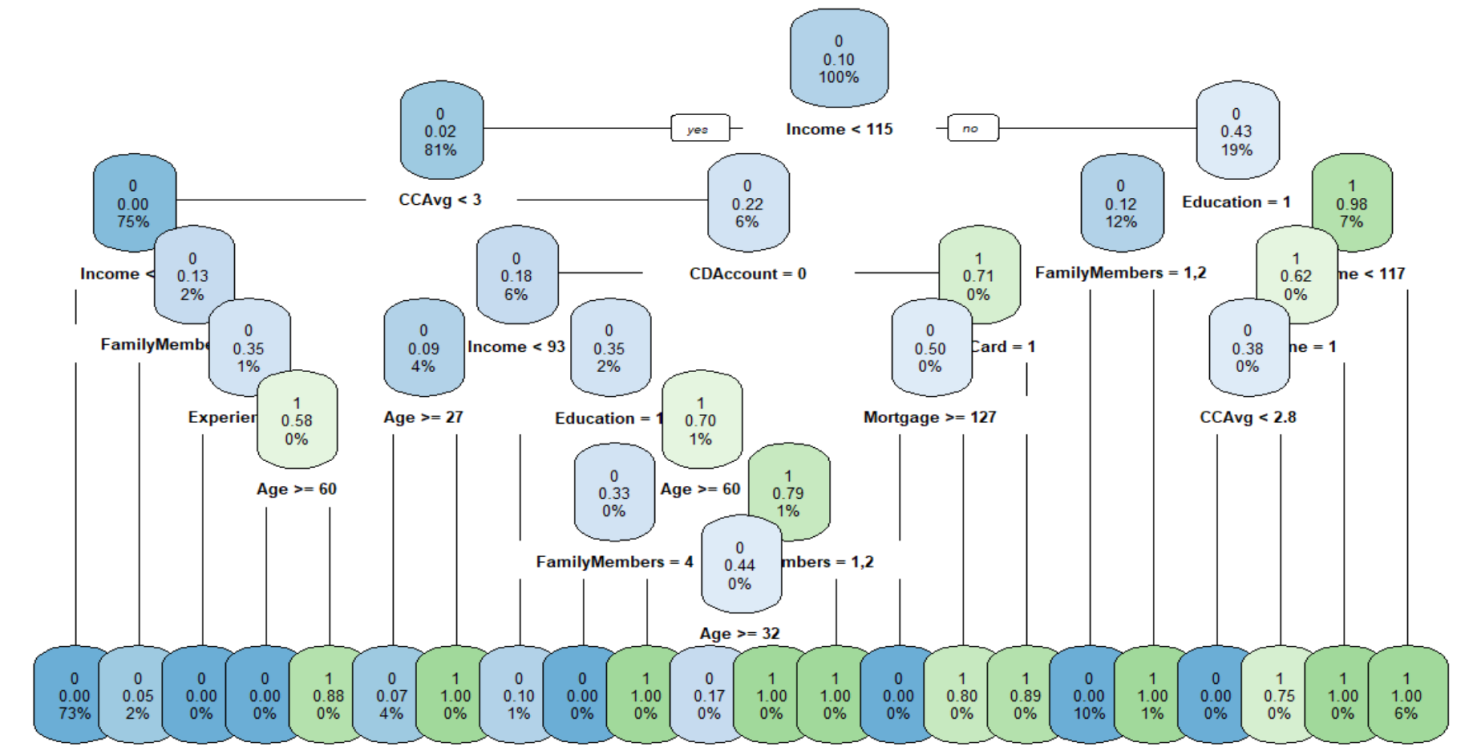
**Pruning the tree using Complexity Parameter**

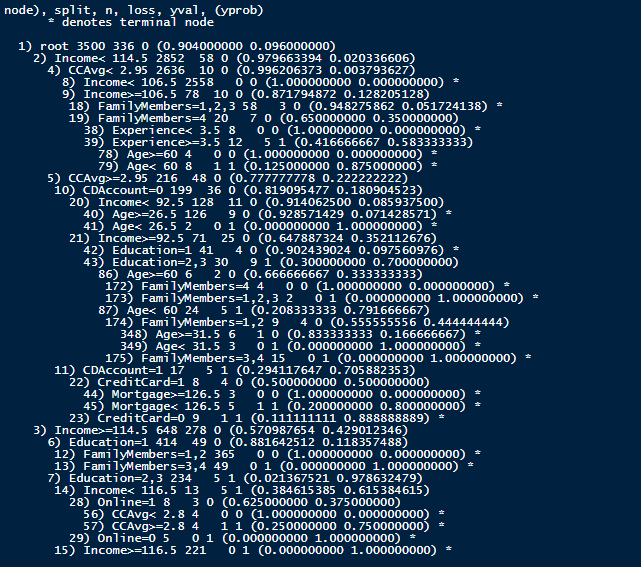
After printing the CP of full tree we observe that cross validation error is minimum at **CP 0.0029762** which is evident from below two plots.

****

****

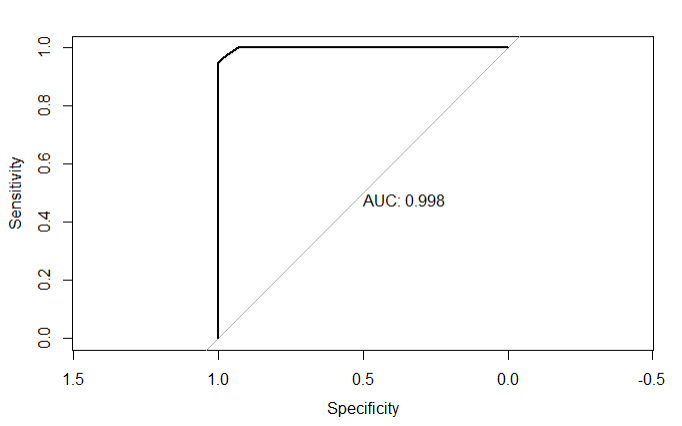
**Building tree using the best CP and plotting the tree**

****

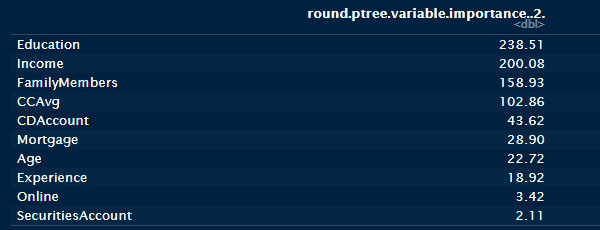
** **

Observation

* Even after pruning the tree using best CP, the model performance on test data is not better than full tree.
* Pruning using CP gives slightly better results ~87% than manual pruning ~83%.
* First split is now happening at whether Income is less than or greater than $ 114.5K
* Second split happens on whether the monthly credit card is less than or greater than 2.95K.
* Even though full tree is giving best results in terms of sensitivity we will opt for pruned tree using best CP. Pruning reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of over fitting. Also the pruned tree using CP has a very high AUC ~**99%**



**Identifying the variable importance for pruned tree using CP**

****

The variable importance coincides with our EDA findings. Factors affecting the model are

1. Education
2. Income
3. Family Members
4. CC Avg
5. CD Account
6. Mortgage

## 3.3 Applying Random Forests (plot the tree)

## 3.3 Interpret the RF model output (with remarks, making it meaningful for everybody)

# Various Model Performance Measures (Test & Train): Confusion Matrix

# Remarks on Model validation exercise (Which model performed the best)